Acta Scientiarum

http://periodicos.uem.br/ojs/acta
ISSN on-line: 1807-8621
Doi: 10.4025/actasciagron.v42i1.43734

A longitudinal study of sweet orange flowering with grouped count data

Idemauro Antonio Rodrigues de Lara[1*], Cesar Augusto Taconeli[2], Rafael de Andrade Moral[3], John Hinde[4], Vanessa Voigt[5] and Sílvia Maria de Freitas[6]

[1]Departamento de Ciências Exatas, Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Av. Pádua Dias, 11, 13418-900, Piracicaba, São Paulo, Brazil. [2]Departamento de Estatística, Universidade Federal do Paraná, Curitiba, Paraná, Brazil. [3]Department of Mathematics and Statistics, Maynooth University, Maynooth, Ireland. [4]National University of Ireland-Galway, School of Mathematics, Statistics and Applied Mathematics, Galway, Ireland. [5]Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, São Paulo, Brazil. [6]Departamento de Estatística e Matemática Aplicada, Universidade Federal do Ceará, Fortaleza, Ceará, Brazil. *Author for correspondence. E-mail: idemauro@usp.br

ABSTRACT. The orange variety “x11”, which is a spontaneous mutant of the sweet orange, has a short juvenile period with early flowering. The data used in this paper are from a randomized design experiment that aimed to assess the plants' flowering characteristics when grafted onto two different varieties of lemon rootstock. The plants were pruned in each of the four seasons, and on each pruning occasion, the number of branches on each plant was counted and classified into four mutually exclusive flowering categories. The data presented large variability and many zeros. The statistical analysis included the use of generalized linear mixed models with a Bayesian approach. The results showed that flowering is not equal over the seasons, i.e., there are significant differences in the classification of the branches across the four seasons and the two varieties, with interactions between seasonal and branch effects.

Keywords: *Citrus*; discrete data; mixed model; Bayesian analysis.

Received on July 17, 2018.
Accepted on October 1, 2019.

Introduction

Citriculture, especially that of the sweet orange (*Citrus sinensis*), is of great importance to Brazil’s economy, and the improvement of citrus production is a constant target for agricultural research. Systematic pruning and flowering encouragement are important for good fruit production. Fruit formation is the end result of a complex chain of events during plant development, in which flowering is a critical step (Goldschimdt & Koch, 1996). Spiegel-Roy and Goldschmidt (1996) note that flowering is strongly influenced by environmental conditions, such as temperature and humidity. This fact was also recorded by other authors, such as Ribeiro, Machado, and Brunini (2006), who studied the environmental conditions of São Paulo under the flowering of citrus, and Nishikawa et al. (2009), who described the correlation between flowering and temperature in citrus trees that have seasonal flowering periodicity.

Additionally, some Citrus genus species have complex biological characteristics, resulting in difficulties for the genetic improvement of flowering and fruiting (Grosser & Gmitter, 1990). However, plants with a short juvenile cycle have excellent potential for use in crop improvement studies. Among the numerous citrus species, orange variety "x11", which is a spontaneous mutant of sweet orange, has a short juvenile period, with early flowering within one or two years of cultivation. This quality makes this variety an excellent choice for functional genomic studies of flowering and fruiting features (Tan & Swain, 2006). According to Pompeu Júnior (1991), the rootstocks also influence citrus production. The robustness of plants against environmental conditions, such as stresses of biotic and abiotic origin, is influenced by the rootstocks (Medina & Machado, 1998).

On this basis, we present an experiment to assess the flowering characteristics of “x11” under two rootstocks, namely, Rangpur lime and Swingle citrumelo, over four seasons. These two rootstocks are the most used in Brazil due to their good production, cold tolerance and resistance to pests (Schäfer, Bastianel, & Dornells, 2001). In addition to the importance of this study from the genetic and agronomic viewpoints, this research also highlights the analysis of a type of response variable commonly observed in plant studies: a longitudinal count of a categorical response, which refers to the type of flower.

The generalized linear model framework (Nelder & Wedderburn, 1972) with extensions for categorial data (Agresti, 2002) can be used to analyze this type of dataset. In some situations, when the sample size is small and sparse, this framework may result in estimation problems, producing nonsensical estimates associated with infinite standard errors. There are some alternatives to handle this shortcoming, such as using bootstrapping (Efron, 1979) or using flexible distributions belonging to the Tweedie family (Bonat et al., 2017). Additionally, in such circumstances, Bayesian modeling can serve as an alternative, providing an informative prior for the parameters associated with the problematic estimates attained under the frequentist approach. However, these methods are rarely used in the analysis of citrus flowering data. In this context, this paper aims to contribute to the analysis of flowering data on the sweet orange variety "x11" using mixed effects models combined with a Bayesian approach.

Material and methods

The data used in this paper refer to an experiment conducted in 2011 in a greenhouse at the Center of Citrus Sylvio Moreira, located in Cordeirópolis City, São Paulo State, Brazil (latitude 22°27'42.6" S, longitude 47°23'57.1" W, altitude 668 m). The main objective of the research was to evaluate the flowering of adult plants of sweet orange on two rootstocks, namely, Rangpur lime (*Citrus limonia* Osbeck) and Swingle citrumelo (*Citrus paradisi* Macf. x *Poncirus trifoliata* (L.) Raf.) over four seasons (spring, summer, autumn, and winter). The study involved 17 similar two-year old plants, cultivated in 20-liter pots under controlled conditions of irrigation and fertilization and arranged in a completely randomized design on the benches of the greenhouse. Of these adult plants, 9 were grafted onto Rangpur lime and 8 onto Swingle citrumelo, although one of these plants subsequently died, leaving only 16 plants for the study.

On the same day, the branches of each plant were nonseverely pruned to synchronize the development of the plants. The flowering evaluation was also carried out on a single day for each of the four seasons. For each plant, the number of new branches were counted that had developed with a terminal flower (category 1), with a lateral flower (category 2), with no flowers (category 3), and with aborted flowers (category 4); the total number of branches was classified into four nominal flowering categories.

In this study, despite the small sample size (true replication from 16 plants) there are large numbers of observations (numbers of branches) per plant and per season, in a grouped data structure, yielding a rich dataset for analysis.

The analysis of data of this nature is not straightforward, since the classical standard analysis of variance cannot be applied. A Poisson regression for a discrete count response can be used. Additionally, to model the longitudinal structure of the data, with each plant being observed on four separate occasions, it is possible to fit a Poisson regression model using the generalized estimation equation (GEE) methodology proposed by Zeger and Liang (1986) for marginal models. The GEE framework requires only the specification of a variance function for the response variable to obtain reliable parameter estimates, considering a correlation structure. The idea of this procedure is to introduce, in the estimation process a correlation matrix, $R(\alpha)$, where α is a parameter vector that fully characterizes the matrix correlation. This estimation method focuses on regression parameters β, while ϕ, a scale parameter, and α are treated as nuisance parameters. In marginal models, the population mean response is modeled as a function of the covariates considered, which is relevant when there are balanced data and hypotheses aimed at testing the effects of factors on the population mean.

In the present study, however, there is a complication in the data structure that limits the use of marginal models; this complication, called overdispersion, is caused in part by the presence of many "zeros" and the varying frequency categories over the seasons. Due to the grouped structure of the data, the high occurrence of zero count responses and the small sample size, modeling this dataset is challenging, as we need to accommodate extra-variability. In this context, the mixed effect modeling framework is appropriate.

In the generalized linear mixed model, we allow for this dependence through the inclusion of pertinent random effects in the linear predictor (Breslow & Clayton, 1993; Diggle, Heagerty, Liang, & Zeger, 2002; Pinheiro & Bates, 2000). The correlation between repeated observations can be considered to arise from the presence of a latent (random) variable (Diggle et al., 2002) that can then be allowed for in the comparison of the individual response profiles. These models are called subject-specific models. Additionally, we will use both classical and Bayesian approaches to analyze the dataset. Details of the statistical procedures adopted are as follows.

To establish a basic and general notation, let y_{ijkl} denote the count of the l-th response category for the i-th unit at the j-th time (season) under the k-th treatment (rootstock type). Therefore, the vector y_{ijk} = (y_{ijk1}, y_{ijk2}, ..., y_{ijkC})' represents the individual profile of responses across the C categories for unit (plant) i, at time j and treatment group k, i.e., the observed frequencies of each response category. Here, the response variable refers to the classification of branches into mutually exclusive categories; we have y_{ijk1} (the number of branches with a terminal flower), y_{ijk2} (the number of branches with a side flower), y_{ijk3} (the number of branches without a flower) and y_{ijk4} (the number of branches with an aborted flower). The sum of the observed frequencies over the C response categories, $m_{ijk.} = \sum_{l=1}^{C} y_{ijkl}$, gives the (marginal) total number of branches for unit i at time j and treatment group k.

Because the total values m_{ijk} are not fixed, the observed frequencies in each of the C response categories are considered as random count variables that can most simply be assumed to follow a count distribution, such as the Poisson distribution, which is a generalized linear model (GLM). Here, because of the repeated measurements of the plants over four seasons (longitudinal data), we are likely to have correlations among the four response profiles for each plant in addition to the associations among the branch category counts for each plant. It is possible to fit a Poisson regression model to such correlated data using a generalized linear mixed model (GLMM). The GLMM is used to incorporate the correlations between observations and to accommodate possible extra-variability due to the presence of excess zeros and disparate frequencies across categories and treatments. The dependence between observations is due to individual plant differences, and this heterogeneity is captured by the inclusion of appropriate random effects.

In GLMMs, it is assumed that the response variables, conditioned on one or more random effects, are independent random variables that have the structure of a GLM (Molenberghs & Verbeke, 2005). Thus, for our example, we consider $m_{ijk.}$ as random variables, and we have the following:

$$Y_{ijkl} \mid b_i \sim \text{Poisson}\ (\mu_{ijkl}), \qquad (1)$$

where: b_i is a vector of individual level random effects, and

$$E(Y_{ijkl} \mid b_i) = \mu_{ijkl}$$

is the conditional mean of the random variable Y_{ijkl}. This mean is related to a systematic linear predictor through a link function as follows:

$$g[E(Y_{ijkl} \,|b_i)] = g(\mu_{ijkl}) = z_i' b_i + x_i' \beta \qquad (2)$$

where: x_i is the covariate vector associated with the fixed effects, β is the fixed effects parameter vector, and z_i is the vector associated with the random effects vector b_i.

The systematic part of the mixed model given in (2) includes both fixed and random effects, and the random effects vector b_i constitutes a sample from a q-dimensional random variable, with the usual (but not only) choice being a multivariate normal distribution, i.e., $b_i \sim N_q(0, G)$. The objective of the analysis is to estimate the coefficients of the fixed effects, β, the parameters of G, the $q \times q$ variance-covariance matrix of the random effects, and, in some cases, an additional dispersion parameter, ϕ.

There is extensive literature on the estimation of the parameters in the GLMM (2), and it is common to use maximum likelihood (Breslow & Clayton, 1993; Diggle et al., 2002; Pinheiro & Bates, 2000). The likelihood function of the vector of unknown parameters ψ, which includes both β and G, is as follows:

$$L(\psi, \phi, y) = \prod_{i=1}^{16} \int \prod_{j=1}^{4} \prod_{k=1}^{2} f(y_{ijkl} | b_i, \beta, \phi)\ \ f(b_i\,, G) db_i \qquad (3)$$

which is obtained by integrating (marginalizing) the joint distribution of (Y, b) over the random effects b.

The problem with maximizing the function in (3) is the presence of 16 q-dimensional integrals over the random effects b_i. In most cases, these integrals have no exact analytical solution; thus it is necessary to use numerical methods to obtain an approximate solution.

We used the mixed model structure, starting with the maximal linear predictor as follows:

$$\eta_{ijkl} = \alpha_0 + \delta_j \text{season}_j + \tau_k \text{treatment}_k + \kappa_l \text{category}_l + \gamma_{jk} \text{season}_j * \text{treatment}_k + \beta_{jl} \text{season}_j * \text{category}_l + b_{0i} + b_{ij} + b_{ijkl} \qquad (4)$$

where: α_0 is the intercept, δ_j is the effect of the j-th season, τ_k is the effect of the k-th treatment, κ_l is the effect of the l-th category, γ_{jk} is the effect of the interaction between the j-th season and the k-th treatment,

β_{jl} is the effect of the interaction between the j-th season and the l-th category, $b_{0i} \sim N(0, \sigma_p^2)$ is a random effect related to observations of the i-th plant, $b_{ij} \sim N(0, \sigma_{ps}^2)$ is a random effect related to observations of the i-th plant and j-th season, and $b_{ijkl} \sim N(0, \sigma_u^2)$ is an observation-level random effect included to accommodate extra-variability.

To fit the mixed models as in (4) by maximizing the likelihood in (3), we used the glmer function of the lme4 package available in the R software (R Core Team, 2018) to attain the full solution using adaptive Gaussian quadrature to approximate the integral (the default uses only one point, corresponding to the Laplace approximation). From model (4), a set of submodels were tested using likelihood ratio tests, but we kept the same random effects because they reflect the hierarchical structure of the data. We also considered the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the measure of deviance to perform model selection.

Due to the observation of only zeros for the combinations Autumn:lateral flowers and Summer:no flowers, the estimation of the associated parameters (β_{32} and β_{23}, respectively) is problematic and associated with large standard errors. Hence, for the selected model, we carried out Bayesian analysis as an alternative to the previous method of fitting by maximum likelihood. Here, we set up prior distributions for the parameters of the model that, combined with the likelihood function through the Bayes rule, provide posterior distributions for them (Gelman et al., 2014). Our objective with Bayesian analysis, besides presenting a different methodology for this dataset, is to overcome drawbacks related to the classical approach. Specifically, we can pinpoint the unreliable inferences for the parameter estimates related to experimental conditions with zero counts, where confidence intervals and hypothesis tests based on asymptotic normal theory are inadequate and provide inconsistent results.

To perform this Bayesian analysis, we set up noninformative prior distributions for all model parameters except for the two regression parameters related to the experimental conditions without any flowers. Initially, we specified the prior distributions for the variance components (σ_p^2, σ_{ps}^2, and σ_u^2) to be inverse gamma IG (0.0001, 0.0001), and a multivariate normal distribution with a mean vector of zeros and an identity covariance matrix with a precision parameter equal to 10^{10} for the regression coefficients.

For parameters β_{32} and β_{23}, we specified informative prior distributions. Although we have observed neither lateral flowers in autumn nor absence of flowers in summer, there is no reason why these could not happen. Thus, it is plausible that these events could happen in other seasons. The interaction coefficient as follows:

$$\exp(\beta_{32}) = \frac{\mu_{(\text{Autumn:lateral flowers})}/\mu_{(\text{Autumn:terminal flowers})}}{\mu_{(\text{Spring:lateral flowers})}/\mu_{(\text{Spring:terminal flowers})}}$$

corresponds to how many times the ratio between the mean number of side and terminal flowers is greater (or smaller) in autumn than in spring (by taking terminal flowers and spring as baseline categories in our model matrix). The same interpretation applies to the interaction coefficient as follows:

$$\exp(\beta_{23}) = \frac{\mu_{(\text{Summer:no flowers})}/\mu_{(\text{Summer:terminal flowers})}}{\mu_{(\text{Spring:no flowers})}/\mu_{(\text{Spring:terminal flowers})}}$$

which corresponds to how many times the ratio between the number of branches without flowers and terminal flowers is greater (or smaller) in summer than in spring.

Now, to specify a prior distribution, assuming normality, we need two pieces of information, namely, (i) an estimate for $\exp(\beta_{jl})$ and (ii) a lower 5% limit associated with this estimate. We believe that the ratio between the mean number of lateral flowers and that of terminal flowers is five times greater in spring than in autumn. Additionally, we believe that the probability that this ratio is greater than 20 is 0.05. These beliefs translate to the following specifications:

$$\exp(\beta_{32}) = \tfrac{1}{5} \Rightarrow \beta_{32} = \ln\tfrac{1}{5} = -1.61,$$

$$\exp\,[-1.61 - 1.64 \times \text{se}(\beta_{32})] = 0.05 \Rightarrow \text{se} = 0.844$$

Analogously, we believe that the ratio between the number of branches with no flowers and that with side flowers is twice as large in spring than in summer, with a probability of 5% that this ratio is greater than 10, leading to the following:

$\exp(\beta_{23}) = \frac{1}{2} \Rightarrow \beta_{23} = \ln\frac{1}{2} = -0.69,$

$\exp\,[-0.69 - 1.64 \times \mathrm{se}(\beta_{23})] = 0.1 \Rightarrow \mathrm{se} = 0.983$

Hence, the prior distributions reflecting these statements are specified as $\beta_{32} \sim N(-1.61, 0.844^2)$ and $\beta_{23} \sim N(-0.69,\ 0.983^2)$.

Posterior summaries of interest were obtained using the R package MCMCglmm, which implements Markov chain Monte Carlo routines for fitting multiresponse GLMMs (Hadfield, 2010). It requires only the specification of the distribution for random variables and prior distributions. We simulated 500,000 samples for each parameter of interest, discarding the first 1,000 as burn-in samples, and the number of simulated samples is large enough to be a satisfactory effective sample size, even in the presence of autocorrelation in the MCMC samples. Convergence for the posterior distributions was checked through density and time series plots.

From the posterior distributions, we computed point estimates for the coefficients (posterior means) and 95% credible intervals using the highest posterior density method and Bayesian p-values for significance of model terms.

Results and discussion

Initially, an exploratory data analysis was performed to identify possible effects of treatments and season, as well as the presence of potential outliers. The means and variances of the number of branches of each type over the four seasons are presented in Table 1. Note that the variances are much larger than the means, and hence, the Poisson model is not a reasonable assumption for these data.

Table 1. Means and variances of the number of branches according to season and flower condition.

	Terminal flower		Lateral flower		No flowers		Aborted flower	
Season	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
Spring	100.13	294.38	1.75	3.67	10.31	20.89	19.06	28.99
Summer	0.75	2.47	0.25	0.60	0.00	0.00	24.87	55.45
Autumn	56.94	154.99	0.00	0.00	25.19	94.83	5.94	30.19
Winter	19.06	272.46	80.87	790.25	3.06	6.86	0.94	0.33

The following can also be observed from Table 1: there was a large incidence of terminal flowers in spring and autumn, plants with lateral flowers were prevalent in winter, plants without flowers were more common in autumn than in other periods, and plants with aborted flowers were predominant in summer.

These patterns are the same for both types of rootstock. These results suggest that flowering in summer and autumn is less intense. These results agree with the studies presented by Guardiola (1997), who stated that flowering in summer and autumn is less intense than that in spring and winter.

In the raw count data, we observe a large percentage of zeros (28.51%), including two combinations of season and flower type that resulted in only zeros (Autumn: lateral flowers and Summer: no flowers). However, there is no structural reason for this result, i.e., this result is a random phenomenon that can occur in other years and seasons. Additionally, it is clear that the response categories change according to the season (see Figure 1).

The estimates of the selected mixed model are presented in Table 2. In this table, it is possible to see that the main effects of treatment, season and category were significant, as were most components of the interactions between them, except the interaction between autumn and side flower (season 3: category 2) and the interaction between summer and no flower (season 2: category 3). However, the size of these estimates and their associated standard errors highlights a problem: these parameters are attempting to reproduce the zero marginal counts for these categories, which calls into question the reliability of the conclusions based on this model.

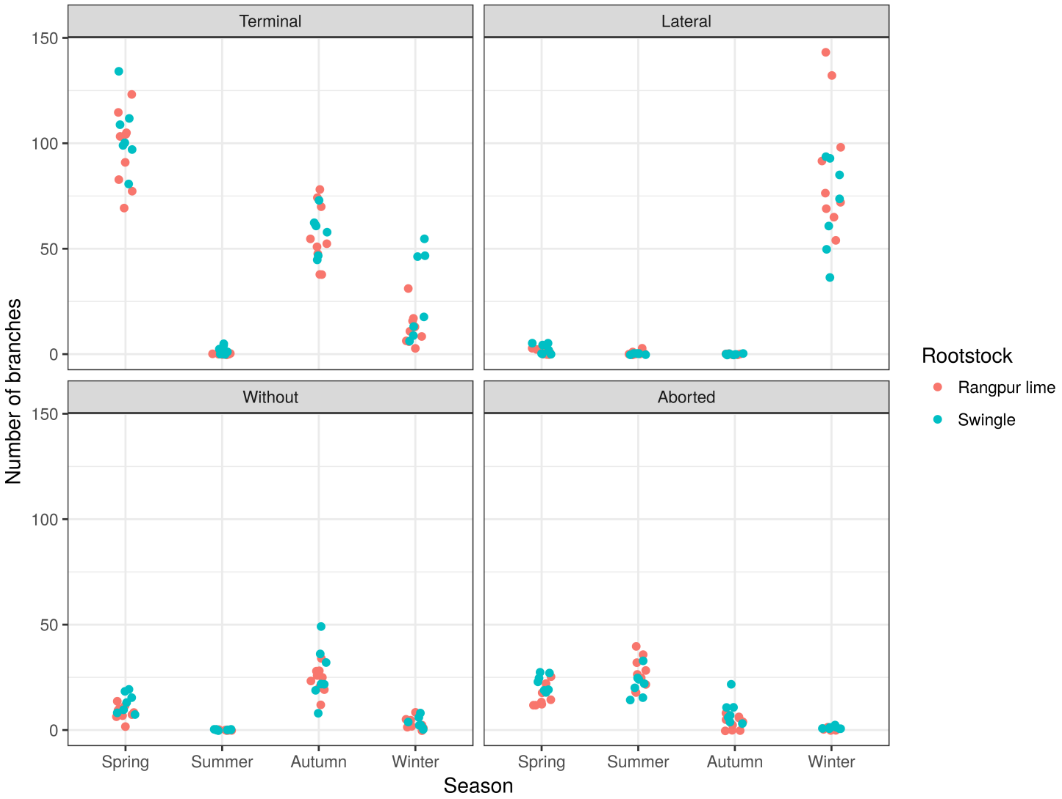

Figure 1. Individual response profiles for the number of branches of each type of treatment over the four seasons.

According to the defined methodology, we first fitted equation (4) and then various submodels to the data retaining the mixed model random effects. We sequentially tested the significance of each fixed factor. By studying the effect of the interaction between treatment and season, it was observed that this effect was not significant (p = 0.3088). However, it was observed that there was an interaction between season and category (p < 0.0001).

Table 2. Parameter estimates and standard errors (se) for the model estimated using the classical approach, and parameter estimates and associated lower and upper 95% credible intervals (l-95% and u-95%, respectively) for the model estimated using the Bayesian approach (Seasons: 1 = spring (baseline), 2 = summer, 3 = autumn, 4 = winter; categories: 1 = terminal flower (baseline), 2 = lateral flower, 3 = no flower, 4 = aborted flower).

Parameters	Classical		Bayesian		
	Estimate	se	Estimate	l-95%	u-95%
(Intercept)	4.50	0.10	4.51	4.30	4.72
Treatment	0.21	0.07	0.20	0.05	0.37
Season 2	-4.96	0.32	-5.06	-5.71	-4.42
Season 3	-0.57	0.13	-0.60	-0.88	-0.33
Season 4	-1.83	0.14	-1.85	-2.15	-1.55
Category 2	-4.11	0.23	-4.22	-4.70	-3.76
Category 3	-2.32	0.15	-2.34	-2.65	-2.04
Category 4	-1.68	0.14	-1.70	-1.99	-1.41
Season 2: category 2	3.03	0.65	3.08	1.73	4.39
Season 3: category 2	-19.27	3989.98	-2.28	-3.58	-1.04
Season 4: category 2	5.69	0.27	5.81	5.27	6.38
Season 2: category 3	-16.71	4059.48	-1.77	-3.33	-0.27
Season 3: category 3	1.48	0.20	1.51	1.09	1.93
Season 4: category 3	0.60	0.25	0.62	0.10	1.13
Season 2: category 4	5.23	0.35	5.32	4.63	6.04
Season 3: category 4	-0.67	0.22	-0.65	-1.09	-0.20
Season 4: category 4	-1.20	0.33	-1.21	-1.88	-0.56
σ_p^2	< 0.0001	0.0028	0.0057	0.0005	0.0236
σ_{ps}^2	< 0.0001	0.0103	0.0048	0.0005	0.0194
σ_u^2	0.1245	0.0201	0.1412	0.0948	0.2013

To account for this problem and explore the reliability of inferences on the other parameters, applied the alternative Bayesian estimation procedure to the selected model. This Bayesian fitted model is summarized in Table 2.

It can be seen that almost all results agree with those provided using the classical (likelihood) approach, except for the two parameters related to the experimental conditions in which we have only zero counts (season 3: category 2 and season 2: category 3) for which we have used specific (slightly) informative priors.

Figure 2 presents diagnostic plots for the coefficients β_{32} (autumn: lateral flowers) and β_{23} (summer: no flowers), which initially presented large standard errors and are associated with potential estimation problems. There is not any evidence of lack of convergence for their posterior distributions. Additionally, we simulated other chains for these parameters to analyze if they also converge (results omitted). Once again, no problems were obtained. Additionally, a similar study of convergence was done for the other parameters and, again, no convergence problems were identified.

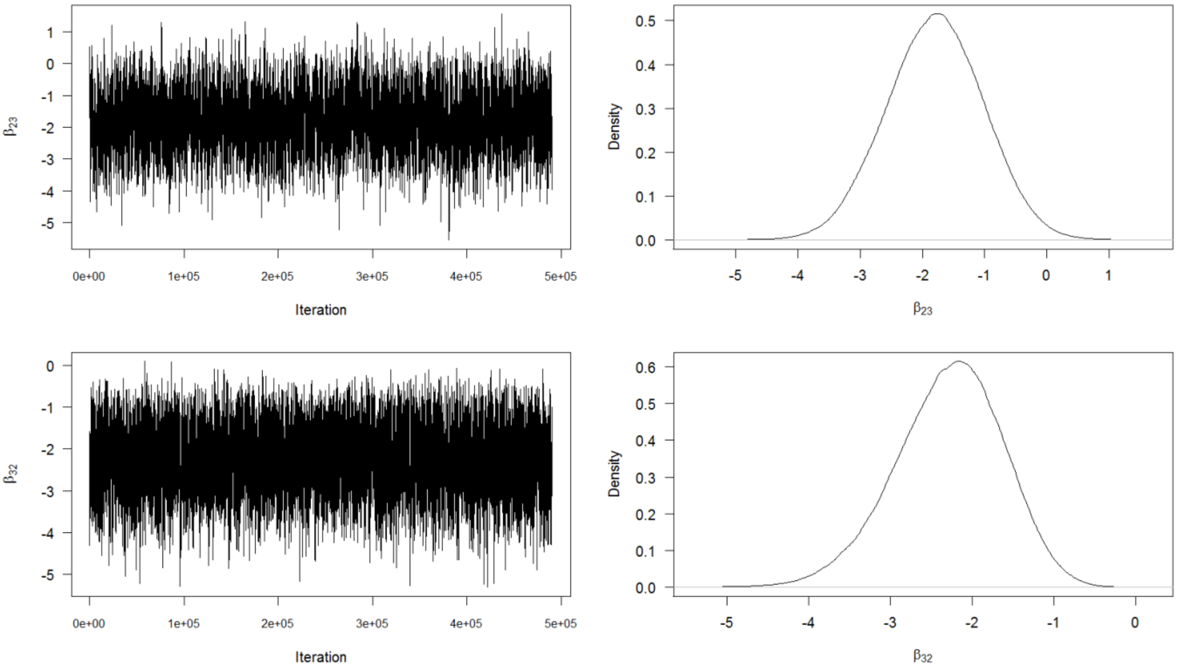

Figure 2. Trace plots and posterior densities for parameters β_{23} and β_{32} for the model estimated using the Bayesian approach.

Figure 3 presents the posterior means for the expected number of flowers, along with credible intervals, for each experimental condition.

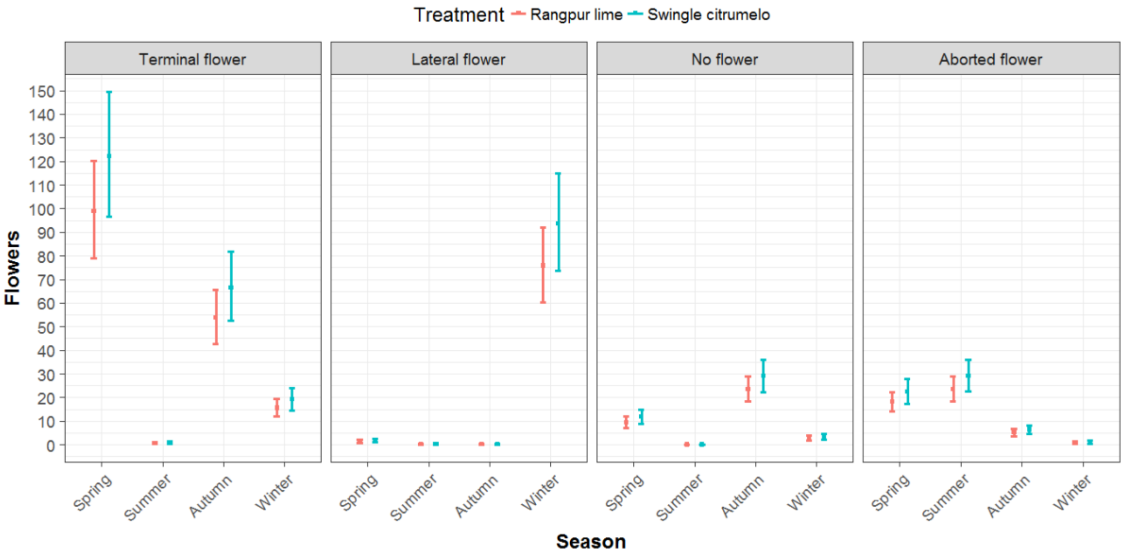

Figure 3. Posterior means and credible intervals for the expected number of flowers under each experimental condition.

Finally, the Bayesian approach allows us to make the following conclusions from the data analysis: i) plants grafted on Rangpur lime and Swingle citrumelo differ statistically in relation to flowering; ii) flowering is not equal across the four seasons, i.e., there are significant differences in the classification of branches over the four seasons; and iii) there is an interaction between season and branch category; i.e., there is a greater abundance of terminal flowers in spring and lateral flowers in winter season; in contrast, there are more aborted flowers in summer and more branches without flowers in autumn (Figure 3).

Therefore, it was found that the diversity and intensity of the types of flowers are related to exogenous factors (captured by random effects) and stress due to the seasons, corroborating the results discussed by Ribeiro et al. (2006).

Conclusion

Our contribution in this work is the use of a methodology for longitudinal count data based on generalized linear mixed models. This methodology is much more appropriate than the traditional techniques based on analysis of variance, in which factorial treatment designs are considered, with time being one of these factors. Additionally, when working with count data, the occurrence of zeros is not uncommon, and this occurrence is a possible cause of overdispersion. In such cases, mixed effects models allow for the accommodation of the extra-variability and the correlations among observations.

Moreover, in our study, the purely classical approach failed to produce reliable standard errors for two parameters of the interaction term. Here, the use of the Bayesian approach allowed for a more stable analysis with more precision and conclusions that corroborated those of the classical model.

Flowering is obviously an important step in the development and production of citrus fruit; however, it is influenced by many endogenous and exogenous factors and selecting an appropriate model can aid in the understanding of the underlying processes. Here, our modeling strategy allowed for the identification of significant differences between the flowering of plants grafted on Rangpur lime and Swingle citrumelo, as well as across seasons, with a predominance of terminal flowers in spring, lateral flowers in winter, aborted flowers in summer, and branches without flowers in autumn.

Acknowledgements

Special thanks to FAPESP (Grant No. 2015/02628-2 and 2018/06288-0).

References

Agresti, A. (2002). *Categorical data analysis*. New Jersey, US: Wiley.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9-25. DOI: 10.2307/2290687

Bonat, W. H., Jorgensen, B., Kokonendji, C. C., Hinde, J., & Demétrio, C. G. B. (2017). Extended Poisson-Tweedie: properties and regression models for count data. *Statistical Modelling*, *18*(1), 24-49. DOI: 10.1177/1471082X17715718

Diggle, P. J., Heagerty, P. J., Liang, K. Y., & Zeger, S. L. (2002). *Analysis of longitudinal data*. New York, US: Oxford University Press.

Efron, B. (1979). Bootstrap Methods: another look at the Jacknife. *The Annals of Statistics*, *7*(1), 1-26.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis*. Boca Raton, US: Chapman & Hall.

Goldschimdt, E. E., & Koch, K. E. (1996). *Photoassimilate distribution in plants and crops: source-sink relationships*. New York, US: Marcel Dekker Inc.

Grosser, J. W., & Gmitter-Junior, F. G. (1990). Protoplast fusion and citrus improvement. *Plant Breeding Reviews*, *8*, 339-374. DOI: 10.1002/9781118061053.ch10

Guardiola, J. L. (1997). Overview of flower bud induction, flowering and fruit set. In S. H. Futch, W. J. Kender, & F. L. Lake Alfred (Eds.), *Citrus flowering and fruit* (p. 5-21). Gainesville, FL: Citrus Research and Education Center.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software*, *33*(2), 1-22. DOI: 10.18637/jss.v033.i02

Medina, C. L., & Machado, E. C. (1998). Trocas gasosas e relações hídricas em laranjeira "Valência" enxertada sobre limoeiro "Cravo" e Trifoliata e submetida à deficiência hídrica. *Bragantia*, *57*(1), 15-22. DOI: 10.1590/S0006-87051998000100002

Molenberghs, G., & Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York, US: Springer Verlag.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A*, *135*(3), 370-384. DOI: 10.2307/2344614

Nishikawa, F., Endo, T., Shimada, T., Fujii, H., Shimizu, T., Omura, M., & Ikoma, Y. (2007). Increased CiFT abundance in the stem correlates with floral induction by low temperature in Satsuma madarin (*Citrus unshiu* Mar.) *Journal of Experimental Botany*, *58*(14), 3915-3927. DOI: 10.1093/jxb/erm246

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects in S and S-PLUS*. New York, US: Springer Verlag.

Pompeu Júnior, J. (1991). Porta enxertos. In O. Rodriguez, F. Viegas, J. Pompeu Júnior, & A. A. Amaro (Eds.), *Citricultura Brasileira* (p. 265-280). Campinas, SP: Fundação Cargill.

R Core Team (2018). *R*: A language and environment for statistical computing. Vienna, AU: R Foundation for Statistical Computing.

Ribeiro, R. V., Machado, E. C., & Brunini, O. (2006). Ocorrência de condições ambientais para indução do florescimento em citros no estado de São Paulo. *Revista Brasileira de Fruticultura*, *28*(2), 247-253. DOI: 10.1590/S0100-29452006000200021

Schäfer, G., Bastianel, M., & Dornells, A. L. C. (2001). Porta-enxertos utilizados na citricultura. *Ciência Rural*, *31*(4), 723-733. DOI: 10.1590/S0103-84782001000400028

Spiegel-Roy, P., & Goldschimdt, E. E. (1996). *The biology of citrus*. Cambridge, UK: Cambridge University Press.

Tan, F. C., Swain, S. M. (2006). Genetics of flower initiation and development in annual and perennial plants. *Physiologia Plantarum*, *128*(1), 8-17. DOI: 10.1111/j.1399-3054.2006.00724.x

Zeger, S. L., & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, *42*(1), 121-130. DOI: 10.2307/2531248

APPENDIX

R CODE TO FIT THE MODELS

```
############################################################################
### Scripts for sweet orange flowering data analysis

### Loading packages
require(MCMCglmm)
require(lme4)

############################################################################
### Classical analysis
fit1 <- glmer(Count ~ treatment + season * category + (1|unity) + (1|plant) + (1|plant:season), data =
final_data, family = poisson)
summary(fit1)

############################################################################
### Bayesian analysis
# Prior specification
I <- diag(17)
eprior <- list(B = list(mu = c(0, 0, 0, 0, 0, 0, 0, 0, 0,- 0.69, 0, -1.61, 0, 0, 0, 0, 0), V = I*c(1e+10, 1e+10, 1e+10,
1e+10, 1e+10, 1e+10, 1e+10, 1e+10, 1e+10, 0.983^2, 1e+10, 0.844^2, 1e+10, 1e+10, 1e+10, 1e+10, 1e+10)),
R=list(V=1, nu=0.002), G=list(G1=list(V=1, nu=0.002), G2 = list(V=1, nu=0.002)))

# Model fitting
fit2 <- MCMCglmm(Count ~ treatment + season * category,
            random= ~ plant + plant:season, data=final_data,
            pl = TRUE,
            prior = eprior,
            family = "poisson", verbose = FALSE,
            nitt = 500000, burnin = 1000, thin = 1, pr = TRUE)
summary(fit2)
############################################################################
```